Bi/CNS 163. Sleep and Dreams

Brief notes on statistics

We tipycally have a situation in science where we have acquired data from two populations and would like to quantitatively compare them. In particular, we wish to know whether the differences that we observe between the two correspond to the parameters or changes that we impose or observe or whether they could simply be the result of random fluctuations. I will describe here one of the simplest ways of comparing two sets of data. I will briefly discuss how to implement this so-called t-test in MATLAB. This requires some strong assumptions about the data. I will very briefly discuss how to test some of those assumptions and some simple suggestion when those constraints are not met. For further information, please refer to any of the basic statistics books in the library, I suggest one particular book at the bottom.

Confidence limits for the difference of means in two normalsize populations: the t test

$x_1, x_2, ..., x_{N_x}$ independent samples from a normalsize distribution with mean μ_x

$y_1, y_2, ..., y_{N_y}$ independent samples from a normalsize distribution with mean μ_y

We will assume common variance: $var(X) = var(Y) = \sigma^2$

We estimate the population means by

$\overline{x} = 1/N_x \sum_{i=1}^{i=N_x} x_i$

$\overline{y} = 1/N_y \sum_{i=1}^{i=N_y} y_i$

and the variance by

$s_x^2 = \frac{\sum_{i=1}^{i=N_x}(x_i - \overline{x})^2}{N_y - 1}$
$s_y^2 = \frac{\sum_{i=1}^{i=N_y}(y_i - \overline{y})^2}{N_y - 1}$
An unbiased estimate of σ^2 based on both samples is
$\hat{\sigma}^2 = \frac{n_x s_x^2 + n_y s_y^2}{n_x + n_y}$
where $n_x = N_x - 1$ and $n_y = N_y - 1$
Clearly, $E(\hat{\sigma}^2) = \sigma^2$
$\overline{x}$ is distributed $\mathcal{N}(\mu_x, \sigma^2/N_x)$
$\overline{y}$ is distributed $\mathcal{N}(\mu_y, \sigma^2/N_y)$
Therefore,
$\overline{x} - \overline{y}$ is distributed $\mathcal{N}(\mu_x - \mu_y, \sigma^2(1/N_x + 1/N_y))$
We can easily construct the statistic
$t = \frac{\overline{x} - \overline{y} - (\mu_x - \mu_y)}{\sqrt{\hat{\sigma}^2(1/N_x + 1/N_y)}}$
which has the Student t-distribution with $N_x + N_y - 2$ degrees of freedom.
The $100(1-\alpha)\%$ confidence limits for $\mu_x - \mu_y$ are therefore given by:
$\overline{x} - \overline{y} + t_\alpha \sqrt{\hat{\sigma}^2(1/N_x + 1/N_y)}$
$\overline{x} - \overline{y} - t_\alpha \sqrt{\hat{\sigma}^2(1/N_x + 1/N_y)}$

t-tests in MATLAB

It is very easy to implement all these equations in MATLAB.
1. Define the variables x and y
E.g.
$x = [3581415161232131514];$
$y = [2423348252912282526];$
(data taken from the Midterm)
2. Write your own code to implement the above equations or simply use MATLAB's ttest2 function. This function takes the two vectors x and y and yields the probability of observing the experimental difference by chance (assuming $\mu_x = \mu_y$ by default). If you are unsure about how to use it or wish to change the value of α type help ttest2 within MATLAB for further information.

Are the data normally distributed?

This is typically a strong assumption, particularly when working with small number of samples and with a problem where we do not understand a priori the distribution of the data. If the sample size is not too small, it is possible to perform a test to address the question of whether the data are distributed normally or not. This is usually referred to in the literature as "goodness of fit" tests. A particularly useful test is the so-called Chi-Square test of goodness-of-fit. It is based on the following theorem:

Theorem: If X_j $(j = 1, ..., n)$ are indipendent and distributed $\mathcal{N}(\prime, \infty)$, and Y_i is a linear orthogonal function of the X_j then $\sum_i Y_i^2$ is chi-square distributed with n defrees of freedom.

Non-parametric statistics

If the sample size is very small or the underlying distribution of the data is unknown, it is possible to use powerful non-parametric tests to address the probability that the differences between the two groups are due to random fluctuations in the data. One such test is the Wilcoxon rank sum test (a variant of which is also called Mann-Whitney test). We are not going to develop here the required equations. The test is based on the idea that, if the numbers are drawn from a single population, then sorting them should not yield any special arrangement of the data. If after, sorting, all the data from one group occur at the beginning and those from the other group at the end, that may reflect some difference between the two groups. These ideas can be easily quantified (see Reference book below).

References

There are plenty of books on statistics and most of them cover the basic situations. One particularly cheap book to buy would be the Dover textbook by Keeping:

Introduction to Statistical Inference E.S.Keeping (1995) Dover Publications, New York

Gabriel Kreiman
gabrielk@caltech.edu